RANDOM-EFFECT MODELS FOR VARIABLE SELECTION

By

Youngjo Lee
Hee-Seok Oh

Technical Report No. 2009-4
March 2009

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065
RANDOM-EFFECT MODELS FOR VARIABLE SELECTION

By

Youngjo Lee
Department of Statistics
Seoul National University

Hee-Seok Oh
Department of Statistics
Seoul National University

Technical Report No. 2009-4
March 2009

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu
Random-Effect Models for Variable Selection

Youngjo Lee
Department of Statistics
Seoul National University
Seoul 151-747, Korea
youngjo@snu.ac.kr

Hee-Seok Oh
Department of Statistics
Seoul National University
Seoul 151-747, Korea
heeseok@stats.snu.ac.kr

Draft: version of March 16, 2009
Abstract

In recent times, simultaneous variable selection and estimation has become an important area of interests. We show how random-effect models enrich this area by introducing a new type of penalty, unbounded at the origin. This leads to the bimodality and singularity of the likelihood surface. We show how the h-likelihood methods overcome such difficulties to allow an oracle variable selection and simultaneously enhance estimation power.

Keywords: Variable selection; LASSO; SCAD; HGLMs; DHGLMs; random-effect models; h-likelihood.

Abbreviated Title: Random-Effect Models for Variable Selection
1 Introduction

Consider the regression model

\[ y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n, \quad (1) \]

where \( \beta \) is a \( d \times 1 \) vector of fixed unknown parameters and \( \varepsilon \)'s are white noises with mean 0 and finite variance \( \phi \). The goal of this study is to select significant variables and to enhance prediction accuracy.

Many variable selection procedures can be described as the penalized least squares (PLS) framework which minimizes

\[ Q_\lambda(\beta, y) = \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|), \quad (2) \]

where \( p_\lambda(\cdot) \) is a penalty function controlling model complexity. With the entropy or \( L_0 \)-penalty, namely, \( p_\lambda(|\beta_j|) = \lambda I(|\beta_j| \neq 0) \), the PLS becomes

\[ \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda |M|, \]

where \( M = \sum_{j=1}^{d} I(|\beta_j| \neq 0) \) denotes the size of the candidate model. This leads to traditional variable selection procedures, which have two fundamental limitations. First, when the number of predictors \( d \) is large, it is computationally infeasible to perform subset selection. Second, subset selection is extremely variable because of its inherent discreteness (Breiman, 1996; Fan and Li, 2001). To overcome these difficulties, several other penalties have been proposed. The \( L_2 \)-penalty yields a ridge regression estimation, but it does not perform variable selection. With the \( L_1 \)-penalty, specifically, the PLS estimator becomes the least absolute shrinkage and selection operator (LASSO), which thresholds predictors with small estimated coefficients (Tibshirani, 1996). LASSO is a popular technique for simultaneous estimation and variable selection, ensuring high prediction accuracy, and enabling the discovery of relevant predictive variables. Donoho and Johnstone (1994) selected significant wavelet bases by thresholding based on an \( L_1 \) penalty. Prediction accuracy can
sometimes be improved by shrinking (Efron and Morris, 1975) or setting some coefficients to zero by thresholding (Donoho and Johnston, 1994). Tibshirani (1996) gave a comprehensive overview of LASSO as a PLS estimation.

LASSO has been criticized on the grounds that a single parameter $\lambda$ is used for both variable selection and shrinkage. It typically ends up selecting a model with too many variables to prevent over shrinkage of the regression coefficients (Radchenko and James, 2008); otherwise, regression coefficients of selected variables are often overshrunken. To overcome this problem, various other penalties have been proposed. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty for oracle variable selection. More recently, Zou (2006) showed that LASSO does not satisfy Fan and Li’s (2001) oracle property, and proposed the adaptive LASSO.

In this paper, we propose the use of random-effect models to generate penalty functions for variable selection. With double hierarchical generalized linear models (DHGLMs) of Lee and Nelder (2006), various scale mixtures can be considered as distributions for $\beta$. To be specific, we suggest the use of a gamma mixture for $\beta$. The resulting family includes the normal-type (bell-shaped), LASSO-type (cusped), and a new (singular) unbounded penalty at the origin. To evaluate the quality of a method, Zou and Hastie (2005) advocated the following two aspects: (1) accuracy of prediction of future data or estimation of parameters, and (2) interpretation of the mode to allow a simple model by identifying non-significant predictors.

Variable selection is particularly important in the interpretation of the model, especially when the true underlying model has a sparse representation. Identifying null predictors enhances the prediction performances of the fitted model. We show that the new unbounded penalty improves variable selection to enhance prediction performance. We also consider an adaptive penalty selection for a better prediction. Zou and Hastie (2005) proposed the elastic net, which improves the prediction of LASSO by sacrificing the interpretation of the model. With numerical studies, we show that the proposed method outperforms LASSO uniformly in both variable selection and prediction.
Park and Casella (2008) studied the fully Bayesian approach for LASSO. They used a conditional Laplace specification to warrant unimodality for fast convergence of the Gibbs sampler and to avoid meaningless point estimates such as mean or median. Until now, finite penalties, leading to unimodal penalized likelihoods (PL), have been studied. Singularities (unbounded likelihood) have been believed to occur when the description of the process generating observation is not adequate. This may be considered the product of unacceptable probability models. (Edwards, 1972). In this paper, we show that the use of singular likelihood (unbounded penalty) at the origin greatly enhances the performance of variable selection. Furthermore, the resulting estimator satisfies the oracle property of Fan and Li (2001). When the underlying model is sparse, it gives, we found, a much higher probability of selecting the true model than all other existing methods. Zou’s (2006) enhancement of LASSO could be viewed as the use of an asymptotically unbounded penalty. Bimodality, we believe, would also be natural in variable selection because we have two choices: either presence or absence of the predictors. For estimation of \( \beta \), Lee and Nelder (1996) proposed to use the mode of the h-likelihood. We show why the anomalies of singularirity and bimodality do not pose a numerical difficulty in finding an appropriate mode.

The use of methods based on random-effect models can enrich the area of simultaneous variable selection and estimation. There are some analogies between the PL and random-effect model approaches. For example, the iterative weighted least squares (IWLS) estimation for random-effect models can be used for the estimation of \( \beta \) (Lee and Nelder, 2006). This turns out to be identical to the use of the local quadratic approximation of Fan and Li (2001) under the model (1), and hence, we can use Hunter and Li’s (2004) modification to allow that delete covariates to be included in the final model. Random-effect models provide new insights and interpretations of IWLS, which explain how the algorithm overcomes the difficulty in non-convex optimization problem in Section 2.2. However, there are differences between the two approaches. In the PL approach, the penalty need not stem from a statistical model. Thus, the tuning parameters cannot be estimated by
model likelihood, and the sandwich estimators are often used for standard error estimators for $\hat{\beta}$ (Fan and Li, 2001). With an underlying random-effect model, likelihood methods can be used to tune parameters, and the inverse of the Hessian matrix from the h-likelihood can be used for standard error estimators of $\hat{\beta}$. In this paper, we follow the PL approach for the tuning parameter to compare various methods under a uniform condition, and highlight that the new unbounded penalty is better than the existing penalties in variable selection.

In Section 2, we introduce a gamma mixture, which can be fitted by an IWLS procedure. Numerical studies and a real data example are provided in Section 3. In Section 4, we show that this model provides random-effect estimators, which satisfy the oracle property. Concluding remarks are offered in Section 5.

2 Gamma mixture model for $\beta$

Suppose that $\beta$ is a random variable such that

$$\beta|u \sim N(0, u\theta),$$

where $\theta = 2\sigma^2$ and $u$ follows the gamma distribution with a parameter $w$ as

$$f_w(u) = (1/w)^{1/w} \frac{1}{\Gamma(1/w)} w^{1/w-1} e^{-u/w}$$

with $E(u) = 1$ and $\text{Var}(u) = w$. This is the DHGLM with the gamma random effect for the conditional variance $\text{Var}(\beta|u) = u\theta$. Throughout the paper, $f_\alpha(\cdot)$ represents the density function with parameter $\alpha$.

Let $K_s(2\sqrt{\alpha/\beta}) = (\alpha/\beta)^{s/2} \int u^{s-1} \exp\{-(\alpha u + \beta/u)du\}/2$ be the Bessel function of the second kind. This model leads to a distribution for $\beta$, characterized by a parameter $w$ as follows

$$f_{w,\sigma^2}(\beta) = \int f_\theta(\beta|u)f_w(u)du = \frac{2^{1/2-1/w}}{\sigma \sqrt{\pi w \Gamma(1/w)}} \left( \frac{\beta}{\sqrt{w\sigma}} \right)^{1/w-1/2} K_{1/w-1/2} \left( \frac{\beta}{\sqrt{w\sigma}} \right).$$

(4)
As an alternative to the $t$-distribution (the inverse-gamma mixture model), this family has been proposed for errors $\varepsilon$ to fit the response of log return (Madan and Seneta, 1990; Castillo and Lee, 2008). Importing random effects in the mean permits modeling of the first two cumulants, while those in the dispersion provides modeling of the third and fourth cumulants. It has the characteristic function
\[
\phi_\beta(t) = (1 + w\sigma^2t^2)^{-1/w},
\]
so that
\[
E(\beta) = 0, \quad E(\beta^2) = \theta \quad \text{and} \quad E(\beta^4) = 3\theta^2(1 + w). \tag{5}
\]
The coefficient of kurtosis is
\[3(1 + w),\]
such that in this family, $w$ can be interpreted as a measure of excessive relative kurtosis of a normal distribution. The effects of raising $w$ are increased probability near the origin as well as increased tail probability at the expense of the probability in the intermediate range (Fung and Seneta, 2007).

In the random-effect model approach, the penalty function $p_\lambda(|\beta_j|)$ stems from a probabilistic model $p_\lambda(|\beta_j|) = -\phi \log f_{w,\sigma^2}(\beta)$. For the estimation of $\beta$, we may use an h-likelihood (here equivalent to the PL),
\[
h^* = h(\beta, w, \sigma^2) = m_1 + m_2, \tag{6}
\]
where $m_1 = \sum_{j=1}^{n} \log f_\phi(y_j|\beta)$, $m_2 = \sum_{j=1}^{d} \log f_{w,\sigma^2}(\beta_j)$, $f_\phi(y_j|\beta)$ is the density function of $y_j|\beta$, which denotes the likelihood of the regression model (1), and $f_{w,\sigma^2}(\beta)$ is the density function of $\beta_j$ in (4). However, we do not use this h-likelihood because the Bessel function is hard to handle and its optimization is not easy caused by the non-convexity of $-f_{w,\sigma^2}(\beta)$. 
2.1 H-likelihood approach

The random-effect model (3) can be written as

$$\beta = \sqrt{\tau e},$$  \hspace{1cm} (7)

where $\tau = 2\sigma^2 u$ and $e \sim N(0, 1)$. With log link, we have an additive model

$$\log (\tau) = \log 2\sigma^2 + v,$$

where $v = \log u$. This leads to an alternative h-likelihood

$$h = m_1 + h_2,$$ \hspace{1cm} (8)

where $h_2 = \sum_{j=1}^{d} \{ \log f_{\sigma^2}(\beta_j | u_j) + \log f_w(\log u_j) \}$,

$$\log f_{\sigma^2}(\beta_j | u_j) = -\frac{1}{2} \left\{ \log (4\pi\sigma^2) + \log u_j + \beta_j^2 / (2\sigma^2 u_j) \right\},$$

$$\log f_w(v_j) = -\log(w)/w - \log \Gamma(1/w) + \log u_j/w - u_j/w,$$

and $f_{\sigma^2}(\beta_j | u_j)$ and $f_w(v_j)$ are the density functions of $\beta_j | u_j$ and $v_j$, respectively. Given $(w, \phi, \sigma^2)$, for the estimation of $\beta$, we may use the profile h-likelihood

$$p(\beta) = h_1|_{u=\hat{u}} + p_2(\beta),$$

where $\hat{u}$ solves $dh/du = 0$ and $p_2(\beta) = h_2|_{u=\hat{u}}$

$$\hat{u} = \hat{u}(\beta) = w \left\{ (2/w - 1) + \kappa \right\} / 4$$ \hspace{1cm} (9)

with $\kappa = \sqrt{4\beta^2/(w\sigma^2) + (2/w - 1)^2}$.

With $w = 0$ and $w = 2$, the maximization of the profile h-likelihood provides the ridge and LASSO estimator, respectively. To be specific, at $w = 0$, the distribution of $u$ becomes degenerated to 1, such that

$$\hat{u}_j = \hat{u}_j(\beta) = \left\{ 2 - w + \sqrt{4w\beta_j^2/\sigma^2 + (2 - w)^2} \right\} / 4 = 1$$
to give $\hat{v}_j = 0$. Note that, by Stirling’s approximation,

$$\log f_{w=0}(\hat{v}_j = 0) = \lim_{w \to 0} \left\{ -\log(w)/w - \log \Gamma(1/w) - 1/w \right\} = \log(2\pi)/2.$$ 

Thus, the resulting penalty becomes the $L_2$-penalty as

$$p_\lambda(||\beta_j||) = -\phi h_2|_{u=1} \propto \phi/(2\theta) \sum_{j=1}^d \beta_j^2.$$

With $w = 2$, $\hat{u}_j = \hat{u}(\beta) = |\beta_j|/\sqrt{\theta}$, so that

$$p_\lambda(||\beta_j||) = -\phi h_2|_{u=\hat{u}} \propto (\phi/\sqrt{\theta}) \sum_{j=1}^d |\beta_j|$$

becomes the $L_1$-penalty.

Given $\phi = \sigma^2 = 1$, Figure 1 shows penalties $-\phi p_2(\beta)$ at various values of $w = 0$, $2$, and $30$. It alters the form from a quadratic shape ($w = 0$) for ridge regressions to a cusped form ($w = 2$) for LASSO and then to an unbounded form ($w > 2$) at the origin. In cases of $w > 2$, it allows an infinite gain at zero. Bell-shaped penalties have been proposed for aggregation (smoothing; Wahba, 1990) and better prediction (Efron and Morris, 1975) of $L_2$-penalty, and cusped ones for simultaneous variable selection and estimation of LASSO (Tibshirani, 1996) and SCAD (Fan, 1997), denoising of $w, \sigma^2$ (soft thresholding; Donoho and Johnstone, 1994). Until now, only a finite penalty has been investigated. In this paper, we illustrate the advantage of using an unbounded penalty to enhance a variable selection. Singularities in Lasso and SCAD mean that their derivatives are not defined at the origin. Given $\lambda$, however, both penalties satisfy that $p_\lambda(0) < \infty$ and $|p_\lambda'(0)| < \infty$, while the new unbounded penalty has $p_\lambda(0) = \infty$ and $|p_\lambda'(0)| = \infty$. The singular penalty $p_\lambda(||\beta_j||) = \lambda|\beta|^p$ at the origin for $0 < p < 1$ has also been considered. However, it has finite penalty even though it is not differentiable at the origin. It can be shown that $p_2(||\beta||) \propto -\lambda/||\beta||$ near the origin.

The Laplace approximation often gives satisfactorily accurate approximation (Lee and Nelder,
Castillo and Lee (2008) derived the Laplace approximation to $h$, which becomes

$$
\ell = m_1 - \log(w)/w - \log(\Gamma(1/w)) - \log(4\pi\sigma^2)/2
$$

$$
\quad - \frac{1}{2} \left\{ \kappa + \log\left(\frac{\kappa}{2}\right) - \frac{2}{w - 1} \log\left(w\left(\frac{\kappa + 2/w - 1}{4}\right)\right) + \log 2\pi \right\}. 
$$

(10)

One good thing is that Bessel function is not present in h-likelihoods $h$ and $\ell$. In DHGLMs, we could use $\ell$ to have the ML estimators for $\phi$, $\theta$ and $w$. However, in the PL approach, because the true model is the regression model (1), $\theta$ and $w$ are not model parameters, and are associated with the tuning parameters in the penalty function. Therefore, it is not possible to have ML estimation for these parameters in the PL approach.

### 2.2 Fitting algorithm

To gain more insight into the content, we assume that the columns of $X$ are orthonormal. Let $z = X^Ty$ and $\hat{y} = XX^Ty$. The PLS (2) can be expressed as

$$
Q_\lambda(\beta, y) = \frac{1}{2} \sum_{i=1}^n (y_i - \hat{x}_i^Tz)^2 + \sum_{j=1}^d Q_\lambda(\beta_j, z_j),
$$
where
\[ Q_\lambda(\beta_j, z_j) = \frac{1}{2} (z_j - \beta_j)^2 + p_\lambda(|\beta_j|). \] (11)

If \( \beta_j \) is symmetric around zero, we have \( p_\lambda(|\beta_j|) = -\phi p_2(|\beta_j|) \). Thus, minimizing the PLS, which is equivalent to maximizing the h-likelihood, reduces the component optimization problem for the model

\[ z_j = \beta_j + e_j, \] (12)

where \( \text{Var}(e_j) = \phi \). For the simplicity of notation, we will suppress the subscript when deemed unnecessary.

Using the results from Lee and Nelder (2006), the mode of \( \beta \) from \( h \) can be obtained by IWLS

\[ \hat{\beta} = (1^T_2 \Sigma^{-1} 1_2)^{-1} 1^T_2 \Sigma^{-1} z_a = z/[1 + \lambda/\hat{u}], \] (13)

where \( \hat{u} \) is defined in (9), \( \Sigma = \text{diag}(\phi, \theta u^T) \) and \( \lambda = \phi/\theta \). This is the least squares estimator for the augmented linear model

\[ z_a = 1_2 \beta + r, \]

where \( z_a = (z, 0)^T, r = (r_1, r_2)^T, 1_d \) is the \( d \)-dimensional vector of ones, \( r_1 = \sqrt{\phi} e_1 \), and \( r_2 = \sqrt{\theta u} e_2 \), and \( e_i \) are i.i.d. random variables from \( \text{N}(0, 1) \).

Note that \( \theta u_i = (a \theta)(u_i/a) \) for all \( a > 0 \). Therefore, \( \theta \) and \( u_i \) are not separately identifiable. Thus, in random-effect models, we take a parameterization that \( E(u_i) = 1 \) for all \( w \). This imposes a constraint on random-effect estimates \( \sum_{j=1}^d \hat{u}_i/d = 1 \). The difference between parameterizations in random-effect models and the PL depends on \( w \). For comparison purposes, with the PL approach, we follow the convention that \( \lambda = \phi/\theta \). Under the parameterization in the PL approach, it may not hold that \( \sum_{j=1}^d \hat{u}_i/d = 1 \). Whatever the parameterization, the proposed algorithm has a common fit for \( \theta u_i \). From (11) and (8), we have

\[ 0 = \partial(-\phi h/\partial \beta) = \beta[1 + \lambda/\hat{u}] - z = \partial Q_\lambda/\partial \beta = sign(\beta)\{|\beta| + p'_\lambda(|\beta|)\} - z. \]
Thus,
\[ \hat{u} = \lambda |\beta| / p_\lambda'(|\beta|) \] (14)
and
\[ \hat{\beta} = z / (1 + p_\lambda'(|\beta|) / |\beta|). \]

From the above algorithm, we could obtain LASSO, SCAD, and adaptive LASSO including the proposed method. In other words, estimates of most existing variable selection methods can be obtained by using different random-effect estimates \( \hat{u} \) in the IWLS of (13). More specifically, the choice of \( \hat{u} = |\beta| \) provides the LASSO solution for the \( L_1 \)-penalty \( p_\lambda(|\beta|) = \lambda |\beta| \). The adaptive LASSO solution of the penalty \( p_\lambda(|\beta|) = 2\lambda |\beta| / |z| \) in Zou (2006) can be obtained by \( \hat{u} = |\beta| |z| / 2 \), where \( z = \hat{\beta}_{\text{ols}} \). SCAD is implemented by setting the random effect estimate as
\[ \hat{u} = |\beta|/\left\{ I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)}{(a-1)\lambda} I(|\beta| > \lambda) \right\}, \]
for some \( a > 2 \). Consider two IWLS algorithms, maximizing \( h \) and \( \ell \). For the former, we use \( \hat{u} \) in (9), while for the latter, we employ
\[ \hat{u} = \frac{w\{\kappa^2(\kappa + 2/w - 1)\}}{4\{\kappa^2 + (\kappa + 2/w - 1)\}}. \]

We have found that both give similar results; therefore, in this paper, we present only the h-likelihood (HL) method maximizing \( h \). We set \( w = 30 \) for the HL method.

Figure 2 provides the solution paths of LASSO, SCAD, adaptive LASSO, and HL (with \( w = 30 \)) methods when the tuning parameter is set to \( \lambda = 2 \). Remember that the ridge and LASSO are the HL with \( w = 0 \) and \( w = 2 \), respectively. In Figure 2, the solutions of HL are closest to the adaptive LASSO among the three methods.

When \( X \) is not orthogonal, we can use an IWLS from Lee and Nelder (2006)
\[ (X^T X + W_\lambda) \hat{\beta} = X^T Y, \]
where $W_\lambda = \text{diag}(\lambda/\hat{u}_i)$ and $\hat{u}_k = \lambda|\beta_k|/p'_\lambda(|\beta_k|)$. This is identical to Fan and Li’s (2001) IWLS based on local quadratic approximation. From Lee and Nelder (1996, 2006), as a variance estimator for $\hat{\beta}$ we use

$$\phi(X^T X + W_\lambda)^{-1}. \quad (15)$$

When $\hat{u} = 1$, this algorithm gives the ridge estimator and its variance, and when $\lambda = \infty$, the OLS estimator and its variance. When $\hat{u} = 0$, $\hat{\beta} = 0$ with $\text{Var}(\hat{\beta}) = 0$. This offers a new insight
into thresholding through the IWLS procedure. We can allow thresholding, by simply taking null random-effect estimator \( \hat{u} = 0 \). However, when \( \hat{u} = 0 \), \( W_\lambda \) is not defined, since the corresponding diagonal element \( 1/\hat{u} \) is undefined. Therefore, we should delete the corresponding predictors. This causes an algorithmatic difficulty, and hence, we employ a perturbed random-effect estimate \( \hat{u}_{\delta,k} = \lambda(|\beta_k| + \delta)/p_\lambda'(|\beta_k|) \) for a small positive \( \delta = 10^{-8} \). Then, \( W_{\lambda,\delta} = \text{diag}(\lambda/\hat{u}_{\delta,i}) \) is always defined. As long as \( \delta \) is small, the diagonal elements of \( W_{\lambda,\delta} \) are close to those of \( W_\lambda \). Therefore, the solutions of the IWLS, \((X^TX + W_{\lambda,\delta})\hat{\beta} = X^TY\), are nearly identical to those of the original IWLS (15). Note that this algorithm is identical to Hunter and Li’s (2008) one for the improvement of local quadratic approximation. In this paper, we report \( \hat{\beta} = 0 \) and zero standard error estimates if all printed eight decimals are zero.

Note here that the proposed penalty \( p_\lambda(|\beta_j|) \) is non-convex. However, the model for \( p_\lambda(|\beta_j|) \) can be written as hierarchically (i) \( \beta_j | u_j \) is normal and (ii) \( u_j \) is gamma; both are convex GLM optimization. Thus, the proposed IWLS algorithm overcomes difficulties of a non-convex optimization by solving two-interlinked convex optimations (Lee et al., 2006).

3 Numerical studies

To assess the empirical performance of various methods, we conducted simulation studies for the regression model (1). We consider various examples below, including cases of aggregation and spare situations with or without grouped variables. Following Tibshirani (1996), Fan and Li (2001) and Zou (2006), we take \( \phi^{1/2} = 3 \) in all the examples, but in Example 3, \( \phi^{1/2} = 2 \). The first three examples have the sample size \( n = 40 \), and the last two have \( n = 100 \).

• Example 1. A few large effects with \( \beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T \): This case has been studied by Tibshirani (1996), Fan and Li (2001) and Zou (2006). The predictors \( x_i \) are i.i.d standard normal vectors. The correlation between \( x_i \) and \( x_j \) is \( \rho|i-j| \) with \( \rho = 0.5 \). Here, the signal-
to-noise ratio is approximately 5.7.

- Example 2. Many small effects with $\beta = (0.85, 0.85, \cdots, 0.85)^T$: The rest of the settings are the same as in Example 1. It has been studied by Tibshirani (1996) and Zou (2006). Here, the signal-to-noise ratio is approximately 1.8. The ridge regression is expected perform well.

- Example 3. Single large effect with $\beta = (5, 0, 0, 0, 0, 0, 0, 0)^T$: The rest of the settings are the same as in Example 1. This case has been studied by Tibshirani (1996). It represents a typical case in which the significant predictors are very sparse. Here, the signal-to-noise ratio is approximately 7. Example 1 might represent a middle case between Example 2 of aggregation and Example 3 of a very sparse case.

- Example 4. Inconsistent Lasso path with $\beta = (5.6, 5.6, 5.6, 0)$: The predictor variables $x_i$ are i.i.d. $N(0, C)$, where

$$C = \begin{pmatrix}
1 & \rho_1 & \rho_1 & \rho_2 \\
\rho_1 & 1 & \rho_1 & \rho_2 \\
\rho_1 & \rho_1 & 1 & \rho_2 \\
\rho_2 & \rho_2 & \rho_2 & 1
\end{pmatrix}$$

with $\rho_1 = -0.39$ and $\rho_2 = 0.23$. Zou (2006) studied this case and showed that LASSO does not satisfy Fan and Li’s (2001) oracle property.

- Example 5. A few large grouped effects: This example has been studied by Zou and Hastie (2005). This is a relatively large problem with grouped variables. This setting is interesting because only a few grouped variables are significant, such that the variables are sparse in terms of groups, but variables within a group all have the same effects. The true coefficients are

$$\beta = (0, \ldots, 0, 2, \ldots, 2, 0, \ldots, 0, 2, \ldots, 2).$$

The pairwise correlation between $x_i$ and $x_j$ is 0.5 for all $i$ and $j$. 

13
On the basis of 100 simulated data, we compare six methods:

- Ridge of Hoerl and Kennard (1970)
- Lasso of Tibshirani (1996)
- SCAD with $a = 3.7$ of Fan and Li (2001)
- The HL method with $w = 30$
- The adaptive HL method described below

To make comparisons under a uniform condition, we use the same algorithm when choosing the tuning parameter $\lambda$, with the OLS being the initial values for $\beta$. The 10-fold cross-validation was applied. We denote the full data set by $T$, and cross-validation training data and test data set by $T - T^s$ and $T^s$ for $s = 1, 2, \ldots, 10$. For each $\lambda$, we obtain the estimator $\hat{\beta}_\lambda^s$ with the test data set $T^s$ removed. Therefore, we compute the cross-validation prediction mean squares error (PMSE)

$$CV(\lambda) = \frac{1}{n} \sum_{s=1}^{10} \sum_{(y_k, x_k) \in T^s} (y_k - x_k^T \hat{\beta}_\lambda^s)^2.$$ 

We use $\lambda$ that minimizes $CV(\lambda)$. Note that we also tried generalized CV to choose $\lambda$; however, the results were similar to 10-fold CV, and hence we omit the results by generalized CV.

We investigated various values $w$ in the range $10 \leq w \leq 270$, and found that performances are similar; we therefore present the results with $w = 30$. We consider an adaptive HL method that selects $w$ among $w = 0$ (Ridge), 2 (LASSO) and 30 which has the smallest PMSE value. A comprehensive study of adaptive choice of $w$ is left for a future study.

We consider several measures to evaluate the proposed methods. For performance measures for variable selection, we report the proportion of selecting the true model (correctly identifying all null coefficients) and the median number of incorrect 0 (depicting the median of coefficients erroneously
set to 0). In the case of performance measures for estimation, we consider the mean squares error (MSE) and mean absolute error (MAE) of estimators

\[ \text{MSE}(\hat{\beta}) = \frac{1}{\text{no. of coefficients}} \sum_i (\beta_i - \hat{\beta}_i)^2 \]

and

\[ \text{MAE}(\hat{\beta}) = \frac{1}{\text{no. of coefficients}} \sum_i |\beta_i - \hat{\beta}_i| . \]

For performance measures for prediction, we also consider PMSE.

From the results of Tables 1 and 2, we have made the following empirical observations:

1. For variable selection, LASSO and SCAD perform similarly. Adaptive LASSO and the HL with \( w = 30 \) greatly improve other methods for all sparse cases. In Example 2, where many small effects exist, they are the poorest. The HL has the best variable selection in sparse cases.

2. For estimation performance, the HL with fixed \( w = 30 \) outperforms the others in all sparse cases. In Example 2, the ridge works the best.

3. The adaptive HL method greatly enhances estimation performance of the fixed HL method in Example 2. Overall, it has the best estimation performance.

Zou and Hastie (2005) proposed the use of the elastic net, considering a penalty \( p_\lambda(\|\hat{\beta}_j\|) = \lambda\{(1 - \alpha)\|\hat{\beta}_j\| + \alpha\hat{\beta}_j^2\} \). It gives a less sparse solution than LASSO but has better prediction. It greatly improves prediction power by sacrificing the performance of variable selection of the LASSO. The adaptive HL outperforms LASSO uniformly in both variable selection and estimation. Good variable selection enhances estimation performance. Adaptive LASSO also has similar property, but is inferior to LASSO in variable selection of Example 2.

We present the PMSE values over three fixed HL methods, \( w = 0 \) (ridge), \( w = 2 \) (LASSO), \( w = 30 \), and the adaptive HL in Table 3. In all sparse cases, the choice \( w = 30 \) gives the smallest PMSE.
Table 1: Proportion of selecting the true model with the median number of incorrect 0 in parentheses

<table>
<thead>
<tr>
<th>Method</th>
<th>Ridge</th>
<th>Lasso</th>
<th>SCAD</th>
<th>Adaptive Lasso</th>
<th>HL(F)</th>
<th>HL(A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>0.00</td>
<td>0.05</td>
<td>0.04</td>
<td>0.28</td>
<td>0.46</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
</tr>
<tr>
<td>Example 2</td>
<td>1.00</td>
<td>0.52</td>
<td>0.54</td>
<td>0.09</td>
<td>0.07</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.5)</td>
<td>(2.0)</td>
<td>(2.0)</td>
<td>(0.0)</td>
</tr>
<tr>
<td>Example 3</td>
<td>0.00</td>
<td>0.15</td>
<td>0.15</td>
<td>0.50</td>
<td>0.74</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
</tr>
<tr>
<td>Example 4</td>
<td>0.00</td>
<td>0.23</td>
<td>0.28</td>
<td>0.68</td>
<td>0.80</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
</tr>
<tr>
<td>Example 5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.11</td>
<td>0.40</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
</tr>
</tbody>
</table>

Table 2: The medians of MSE (first line) and MAE (second line) values with the median absolute deviation in parentheses.

<table>
<thead>
<tr>
<th>Method</th>
<th>Ridge</th>
<th>Lasso</th>
<th>SCAD</th>
<th>Adaptive Lasso</th>
<th>HL(F)</th>
<th>HL(A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>0.264 (0.123)</td>
<td>0.187 (0.147)</td>
<td>0.187 (0.147)</td>
<td>0.186 (0.173)</td>
<td>0.176 (0.206)</td>
<td>0.179 (0.210)</td>
</tr>
<tr>
<td></td>
<td>0.413 (0.117)</td>
<td>0.291 (0.140)</td>
<td>0.290 (0.142)</td>
<td>0.242 (0.144)</td>
<td>0.231 (0.168)</td>
<td>0.235 (0.140)</td>
</tr>
<tr>
<td>Example 2</td>
<td>0.133 (0.069)</td>
<td>0.267 (0.122)</td>
<td>0.267 (0.112)</td>
<td>0.439 (0.165)</td>
<td>0.458 (0.205)</td>
<td>0.136 (0.082)</td>
</tr>
<tr>
<td></td>
<td>0.292 (0.089)</td>
<td>0.436 (0.111)</td>
<td>0.436 (0.107)</td>
<td>0.564 (0.126)</td>
<td>0.593 (0.168)</td>
<td>0.296 (0.096)</td>
</tr>
<tr>
<td>Example 3</td>
<td>0.139 (0.084)</td>
<td>0.042 (0.041)</td>
<td>0.043 (0.039)</td>
<td>0.025 (0.032)</td>
<td>0.009 (0.012)</td>
<td>0.009 (0.013)</td>
</tr>
<tr>
<td></td>
<td>0.305 (0.090)</td>
<td>0.100 (0.068)</td>
<td>0.099 (0.069)</td>
<td>0.065 (0.065)</td>
<td>0.034 (0.032)</td>
<td>0.034 (0.033)</td>
</tr>
<tr>
<td>Example 4</td>
<td>0.228 (0.225)</td>
<td>0.212 (0.254)</td>
<td>0.201 (0.217)</td>
<td>0.167 (0.190)</td>
<td>0.147 (0.153)</td>
<td>0.147 (0.154)</td>
</tr>
<tr>
<td></td>
<td>0.440 (0.262)</td>
<td>0.391 (0.299)</td>
<td>0.358 (0.266)</td>
<td>0.307 (0.222)</td>
<td>0.275 (0.194)</td>
<td>0.286 (0.189)</td>
</tr>
<tr>
<td>Example 5</td>
<td>0.150 (0.035)</td>
<td>0.124 (0.036)</td>
<td>0.125 (0.038)</td>
<td>0.126 (0.041)</td>
<td>0.119 (0.037)</td>
<td>0.120 (0.042)</td>
</tr>
<tr>
<td></td>
<td>0.304 (0.042)</td>
<td>0.242 (0.041)</td>
<td>0.240 (0.041)</td>
<td>0.219 (0.042)</td>
<td>0.202 (0.042)</td>
<td>0.209 (0.049)</td>
</tr>
</tbody>
</table>

Therefore, the unbounded penalty has a merit in variable selection, estimation and prediction. Overall, the adaptive HL would give the best performance in estimation. Therefore, the fixed HL with \( w = 30 \) is a good choice if the model is sparse. In the absence of this information, we may use the adaptive HL method.

3.1 Real example: prostate cancer data

Tibshirani (1996) applied LASSO to the prostate cancer data of Stamey et al. (1989), which has 9 predictors. The tuning parameter for each method is obtained by the 10-fold CV. The proposed HL
Table 3: The median of PMSE values with the median absolute deviation in parentheses.

<table>
<thead>
<tr>
<th>Method</th>
<th>$w = 0$ (Ridge)</th>
<th>$w = 2$ (Lasso)</th>
<th>$w = 30$</th>
<th>HL (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>10.454 (2.772)</td>
<td>10.021 (2.782)</td>
<td>9.540 (2.628)</td>
<td>9.472 (2.643)</td>
</tr>
<tr>
<td>Example 2</td>
<td>9.842 (2.495)</td>
<td>10.812 (2.994)</td>
<td>10.515 (2.931)</td>
<td>9.707 (2.443)</td>
</tr>
<tr>
<td>Example 3</td>
<td>4.887 (1.330)</td>
<td>4.624 (1.231)</td>
<td>3.854 (1.064)</td>
<td>3.854 (1.064)</td>
</tr>
<tr>
<td>Example 4</td>
<td>9.461 (1.168)</td>
<td>9.430 (1.159)</td>
<td>9.365 (1.116)</td>
<td>9.354 (1.107)</td>
</tr>
<tr>
<td>Example 5</td>
<td>14.170 (2.275)</td>
<td>13.005 (2.380)</td>
<td>12.261 (2.301)</td>
<td>12.154 (1.963)</td>
</tr>
</tbody>
</table>

Table 4: Estimated coefficients of the prostate data with standard errors of estimated coefficients by HL methods in parentheses.

<table>
<thead>
<tr>
<th>Term</th>
<th>OLS</th>
<th>Ridge (w=0)</th>
<th>Lasso (w=2)</th>
<th>SCAD</th>
<th>Adaptive Lasso</th>
<th>HL (w=30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lcavol</td>
<td>0.691 (0.102)</td>
<td>0.476 (0.062)</td>
<td>0.593 (0.063)</td>
<td>0.685</td>
<td>0.674</td>
<td>0.651 (0.084)</td>
</tr>
<tr>
<td>lweight</td>
<td>0.225 (0.083)</td>
<td>0.200 (0.060)</td>
<td>0.149 (0.042)</td>
<td>0.220</td>
<td>0.161</td>
<td>0.191 (0.072)</td>
</tr>
<tr>
<td>age</td>
<td>-0.146 (0.082)</td>
<td>-0.066 (0.060)</td>
<td>0.000 (0.000)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>lbph</td>
<td>0.155 (0.084)</td>
<td>0.112 (0.060)</td>
<td>0.039 (0.018)</td>
<td>0.120</td>
<td>0.003</td>
<td>0.108 (0.063)</td>
</tr>
<tr>
<td>svi</td>
<td>0.317 (0.100)</td>
<td>0.244 (0.062)</td>
<td>0.208 (0.049)</td>
<td>0.288</td>
<td>0.207</td>
<td>0.282 (0.081)</td>
</tr>
<tr>
<td>lcp</td>
<td>-0.147 (0.126)</td>
<td>0.047 (0.063)</td>
<td>0.000 (0.000)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>gleason</td>
<td>0.032 (0.112)</td>
<td>0.054 (0.062)</td>
<td>0.000 (0.000)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>pgg45</td>
<td>0.127 (0.123)</td>
<td>0.077 (0.062)</td>
<td>0.021 (0.011)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000 (0.000)</td>
</tr>
</tbody>
</table>

The method is implemented with $w = 0, 2, 30$. Table 4 shows $\hat{\beta}$ and their standard error estimates from (15). The standard error estimates are reported for the methods, which can be derived from the proposed random-effect model. They lean towards those of the OLS if the estimated coefficients are close to OLSs, and shrink toward zero if the estimated coefficients are close to zero. From Table 4, for the variable selection, the results provided by HL are identical to those of SCAD and adaptive Lasso. LASSO gives a non-zero estimate for the predictor pgg45.

4 Oracle property in variable selection

Suppose that we estimate $\beta$ by maximizing the h-likelihoods $h$, $h^*$, and $\ell$ for the random-effect model (3), but the true model is the fixed regression model (1) with $\beta = (\beta_1, \ldots, \beta_d)^T = (\beta_1^T, \beta_2^T)$, where $\beta_1 = (\beta_{11}, \ldots, \beta_{1s})^T$, $\beta_2 = (\beta_{s+1}, \ldots, \beta_d)^T$ and $\beta_2 = 0$. This true model can be considered
as the submodel of the random-effect model (3) at $\theta = \infty$ (Lee and Nelder, 1996). Thus, we are investigating the behavior of the random-effect estimators from the full (random-effect) model, but the true model is its (fixed-effect) submodel.

Figures 3 and 4 show the likelihood surface of $h$ at various combinations of $(z, \lambda)$. Given $\lambda$, as $z$ (OLS of $\beta$) approaches to zero or given $z$ as $\lambda$ becomes large, there is only one maxima at zero. Thus, the corresponding predictor is not selected in the model. Otherwise, bimodality occurs.

In this case, from the Figures, it can be seen that most often, the likelihood surface appears to support the non-null maximum value (selecting the corresponding predictor as necessary) because a perturbation caused by the singularity at origin is in a negligible region. Thus, this singularity at origin, we found, does not pose a numerical difficulty in finding nonzero local maxima.

To study the asymptotic behavior of estimators as the sample size $n$ increases but the number of predictors $d$ remain fixed, it is instructive to consider a model

$$z_j = \beta_j + e_{nj}$$

where $\text{var}(e_{nj}) = \phi/n$. This leads to the PLS

$$Q_{\lambda_n}(\beta_j, z_j) = \frac{1}{2}(z_j - \beta_j)^2 + p_{\lambda_n}(|\beta_j|),$$

(16)

where $p_{\lambda_n}(|\beta_j|) = p_{\lambda}(|\beta_j|)/n$ and $\lambda_n = \lambda/n$.

Given $\lambda$, the penalty becomes negligible as $n$ becomes larger. For example, in soft thresholding, having $p_{\lambda_n}(|\beta_j|) = \lambda_n|\beta_j|$ and $p'_{\lambda_n}(|\beta_j|) = \lambda_n \text{sign}(|\beta_j|)$, the value of $\lambda$ should increase with the sample size $n$ to prevent the penalty from becoming negligible. To have the oracle property for SCAD, Fan and Li (2001) required $\lambda$ to satisfy $n^{-1/2}\lambda \to \infty$. However, algorithmically, $\lambda$ is often chosen by methods such as cross-validation. Thus, there is no guarantee that the value of $\lambda$ satisfies such a condition. With the use of an unbounded penalty, the penalty cannot be negligible because of $p_\lambda(0) = \infty$. Suppose that we want to have $\hat{\beta} = 0 (\hat{u} = 0)$ if $\beta = 0$. From (14) we have

$$\hat{u} = \lambda_n|\beta|/p'_{\lambda_n}(|\beta|) = \lambda|\beta|/p'_{\lambda}(|\beta|).$$
Figure 3: Likelihood functions of different $\lambda$'s with fixed $z = 10$.

Under the unbounded penalty with $p'_{\lambda_n}(0) = \infty$, $\hat{u}$ can approach to zero even when $\lambda = O(1)$. When $\beta \neq 0$, $\hat{u} = O_p(1)$ from (9). For example, in HL,

$$\hat{\beta} = z_n/[1 + \lambda_n/\hat{u}] = z_n/[1 + \lambda/(n\hat{u})]$$

tends towards $z_n$ and therefore $\hat{\beta} \neq 0$. This is the oracle property we want to possess. With a finite penalty, the oracle property that $\hat{\beta} = 0$ if $\beta = 0$ can be achieved under a stringent condition that $n^{-1/2}\lambda \to \infty$ (Fan and Li, 2001). With $p'_{\lambda}(0) = \infty$, the HL estimators satisfy the oracle property under $\lambda = O(1)$. Adaptive LASSO with $p_\lambda(|\beta_j|) = (2\lambda/|z_j||\beta_j|$ has an asymptotically
unbounded penalty function when the true $\beta_j = 0$ because $z_j$ approaches zero asymptotically. The simulation studies in the previous section show that the unboundedness of penalties is very helpful in identifying null predictors in finite samples, enhancing the estimation of the fitted model.

Let $\nabla m(\hat{\beta}) = \partial m/\partial \beta|_{\beta=\hat{\beta}}$ and $\nabla^2 m(\hat{\beta}) = \partial^2 m/\partial \beta \partial \beta^T|_{\beta=\hat{\beta}}$. For the fixed regression model, $m_1$ in (8) is the log-likelihood, such that

$$I(\beta) = -\nabla^2 m_1(\beta)$$

is the Fisher information matrix. Here $I(\beta_1, 0) = -\nabla^2 m_1(\beta_1, 0)$ is the Fisher information matrix.
knowing $\beta_2 = 0$, and $I_1(\beta_1) = -\nabla^2 m_1(\beta_1)$ is the Fisher information matrix without $\beta_2$. Under the random effect model,

$$ H(\beta) = -\nabla^2 h(\beta) = I(\beta) - \nabla^2 h_2(\beta) $$

is the information matrix from the h-likelihood. Here $H(\beta_1, 0) = -\nabla^2 h(\beta_1, 0)$ is the Fisher information matrix knowing $\beta_2 = 0$, and $H_1(\beta_1) = -\nabla^2 h(\beta_1)$ is the Fisher information matrix without $\beta_2$.

**Theorem 1** Under the assumptions of Fan and Li (2001), we obtain the following results:

1. The h-likelihood estimator $\hat{\beta}$ is root-$n$ consistent, and

2. If $p_\lambda'(0) = \infty$ $\hat{\beta}$ satisfies the oracle property, namely,

   (a) Sparsity: $\hat{\beta}_2 = 0$

   (b) Asymptotic normality:

   $$ \sqrt{n}\{H_1(\beta_1)\}[\hat{\beta}_1 - \beta_1 - \nabla^2 h(\beta_1)^{-1}\nabla h_2(\beta_1)] \rightarrow N\{0, I_1(\beta_1)\} $$

   in distribution.

Proof: Because $p_{\lambda_n}(|\beta_j|) = p_{\lambda}(|\beta_j|)/n$ when $\lambda = O(1) \max\{p_{\lambda_n}'(\beta_j) : \beta_j \neq 0 \text{ for } j = 1, ..., s\} \rightarrow 0$. From Theorem 1 of Fan and Li (2001), it can be shown that the h-likelihood estimator converges at the rate $O_p(n^{-1/2})$, because $a_n = \max\{p_{\lambda_n}'(\beta_j)/n : \beta_j \neq 0\} = O(n^{-1})$. This proves the result (1).

We can show that $p_\lambda'(0) = \infty$, regardless of using $h$, $h^*$ and $\ell$. Thus, we have

$$ \lim_{n \rightarrow \infty} \liminf_{\beta_j \rightarrow 0} p_{\lambda_n}'(\beta_j) > 0. $$

When $p_\lambda'(0) = \infty$, for $j = s + 1, ..., d$, we obtain

$$ \frac{\partial h}{\partial \beta_j} = n\{-p_{\lambda_n}'(\beta_j)\text{sgn}(\beta_j) + O_p(1/n^{1/2})\} $$

$$ = n\{-(1 - 2/w)/(n\beta_j) + O_p(1/n^{1/2})\}. $$
Thus, the sign of the score above near the origin $\beta_j = 0$, for example, $|\beta_j| < 1/n$, is totally determined by $\beta_j$. Hence, by using this result and Fan and Li’s (2001) proof, we can show the oracle property of the h-likelihood estimators. □

Because the true model is a submodel of the assumed random-effect model, the asymptotic covariance matrix of $\hat{\beta}_1$ is $H_1(\beta_1)^{-1}$ in (15). The h-likelihood estimator $\hat{\beta}_1$ is an oracle estimators, performing as well as if $\beta_2 = 0$ known if $w \geq 2$.

With an unbounded penalty satisfying $p'_\lambda(0) = \infty$, we immediately see that they satisfy Fan and Li’s (2001) sparsity condition, but not the continuity condition. In solution paths of Figure 2, we fix $\lambda$. However, as $z$ increases, the estimate of $\theta$ may increase because $\text{var}(\beta) = \theta$. With an estimation of tuning parameter $\lambda$, the continuity condition would become tricky; this become more subtle in adaptive HL method. The numerical study in the previous section shows that the discontinuity condition of Fan and Li (2001) may not deteriorate the performance of the HL method. Throughout numerical studies we have not encountered any numerical difficulty caused by starting values.

5 Conclusion

In this paper, we study the new unbounded penalty that has a merit in variable selection, enhancing estimation and prediction. The h-likelihood and penalized likelihood approaches unified various fields of statistics (Lee et al., 2006; Fan and Li, 2006), so that further unification is possible by combining extensions of the both the approaches. Thus, our methods can be extended in various ways. By using the inverse-gamma mixture for errors $\varepsilon$, it is possible to perform a robust regression (Lee et al., 2006). Various non-normal models can be considered via DHGLMs (Lee and Nelder, 2006) and frailty models (Lee et al., 2006).
Acknowledgement

This work was prepared while Youngjo Lee (YL) was visiting the Department of Statistics of Stanford University. YL would like to thank to Professors Tze Lai, Trevor Hastie, Ji Zhu, Johan Lim and Meeyoung Park for their helpful comments.

References


